Systematic Mapping Study on Use of Pre-Trained Open Machine Learning Models

Riku Alho, Mikko Raatikainen, Jukka K. Nurminen, Lalli Myllyaho and Lucy Ellen Lwakatare*

Department of Computer Science, PL 68 (Pietari Kalmin katu 5), University of Helsinki, Finland

Abstract

Accurate understanding of pre-trained open source machine learning models, and their frameworks, and datasets use can help software engineers simplify, reduce costs, and improve the quality of application development to different domains. This paper investigates how pre-trained Open Machine Learning (ML) models, and their frameworks, and datasets are shared and used in different domains. A systematic mapping study is used to identify published studies. Statistical and qualitative results are formed for 499 studies which provide sufficient information regarding the use of open source pre-trained models, frameworks, and datasets. Based on a relatively large sample, the reviewed 499 studies provide a listing of Open ML models, frameworks, and datasets used in research as well as their relative popularity. The selected studies consisted of a large number of different domains, which saw benefits ranging from minor decline to moderate improvement when compared to the previously used state of the art machine learning methods. Most of the models in studies were used under the TensorFlow framework with ImageNet as the dataset. The majority of studies were made in laboratory environments. Pre-trained Open ML models show positive promise for improvement in machine learning. Additional diversity of available open source models pre-trained with different datasets would improve this effect. More comparable studies are needed, especially from the industry, that use and apply open source machine learning, which report their context, methodology, and performance comprehensively.

Keywords

Machine Learning, Open Source, Reuse, Systematic Literature Review, Systematic Mapping Study

1. Introduction

Although there are important differences, machine learning system developers can learn a lot from traditional software engineering [1]. Both begin their task by familiarizing themselves with the problem domain. Software engineers explore existing and similar solutions, software, and databases, whereas machine learning engineers explore available machine learning options, models, frameworks, and datasets for the problem domain.

Traditionally many technologies related to machine learning have been hidden behind technology industry walls. Not until recently, we have seen a large and systematic introduction of multiple, new open source machine learning technologies: Such shared off-the-shelf pre-trained open source *machine learning models*, *frameworks*, and *datasets* can provide competitive state of the art capabilities in terms of performance, cost-effectiveness, and adaptability in different application domains when applied to new machine learning problems. This may provide affordable new avenues for software engineers, researchers, students, businesses, and private enthusiasts to help reap the benefits of available data without requiring them to invest work on reinventing the wheel [2].

The goal of this study is to assess the shared usage, adoptability, and evolvability of pre-trained open source machine learning models in different application domains. The study is carried out as a systematic mapping study [3], now an established research method in computer science to systematically collect an overview of research state-of-the-art.

This review paper is structured as follows. First, Section 2 introduces to the terminology and background of this study. In Section 3, the research questions and applied research method are presented. The results are introduced and analysed in Section 4. The analysed results and findings are discussed on the Section 5. Finally, Section 6 concludes the paper.

2. Background

To properly understand the terminology and background for this study, we briefly describe the different terms and fields related to it.

The Annual Symposium of Computer Science, June 13-14, 2023

*Corresponding author.

riku.alho@helsinki.fi (R. Alho); mikko.raatikainen@helsinki.fi (M. Raatikainen); jukka.k.nurminen@helsinki.fi (J. K. Nurminen); lalli.myllyaho@helsinki.fi (L. Myllyaho); lucy.lwakatare@helsinki.fi (L. E. Lwakatare)

2.1. Basic concepts

Artificial Intelligence (AI) consists of all technical aspects that aim to get computers to imitate intelligent behaviour observed in humans [4]. This includes machine learning, natural language processing (NLP), language synthesis, computer vision, robotics, sensor analysis, optimization, and simulation.

A subset of AI is *Machine Learning* (ML), which consists of techniques that enable computers to change their functionality based on given information (e.g., sensor data), thus improving their behaviour to achieve the goal [5]. ML techniques include decision trees, neural networks, support vector machines, and many more.

Neural Networks (NNs) are a part of ML. They are computer programs inspired by biological neural network processes [6]. These consist of perceptrons, convolutional neural networks, recurrent neural networks, Boltzmann machines, deep neural networks, and many more. Basic NNs with one to a few layers of neurons usually require user assistance in forming classification classes. *Deep Neural Networks* (DNNs) are under the NN category. They are neural networks, which consist of multiple layers providing them the ability to form new classification classes.

Machine learning can be categorized into *supervised, unsupervised learning, and reinforcement learning* [7]. Supervised learning utilizes training data for classification and regression. Unsupervised learning constructs predictions of classification based on the given input data. Reinforcement learning uses trial and error based on an oracle, such as a repeatable simulation or a game, to find the optimal outputs.

Open source refers to a computer program for which the source code is available to the general public for use or modification from its original design [8]. Open source code is a collaborative effort where programmers improve upon the source code and share the changes within the community. Code is released with a license specifying the conditions under which others may download, use, modify, and publish their versions to the community. This view to open source is not restricted to any particular license.

2.2. Open ML Models

ML models are computer programs or components that have formed statistical and mathematical insights from data, such as a trained neural network. Although ML models usually refer to something already trained, they have also been used to refer to untrained, manually tuned, or default-valued programs. Statistical and mathematical insights can be formed with machine learning or manual tuning.

Training an effective ML model requires large amounts of training or input data, which can be challenging to acquire. *Transfer learning* can be used to mitigate these challenges [9]. Using a pre-trained model taught with large amounts of data that even slightly overlaps the targeted domain may achieve comparable or even better results than using only small datasets available to the targeted domain in question.

Specifically, in this paper, we use the term *Open ML model* for pre-trained open-source machine learning models that can be reused as such or after retraining as components in other systems.

2.3. Frameworks

The majority of Open ML models are used inside dedicated *frameworks*, such as Caffe [10], Keras[11], Weka [12], PyTorch[13], TensorFlow[14] or MatConvNet [15]. Frameworks work as the interface between an ML model, users, and hardware and can, thus, affect how hardware calculations and values are given to the model during training and use.

2.4. Datasets

Most off-the-shelf open ML models offered by different frameworks are made available pre-trained under a certain *dataset.* There are many different datasets with varying scales ranging from entries of tens of thousands to a billion. Datasets, such as ImageNet [16], Places365 [17], CIFAR [18] and Pascal VOC [19], are offered in pre-trained open ML models. The benefit of pre-trained models is that training a model comprehensively on a large dataset takes a very long time and a lot of computing resources and data. For instance, one study [20] reports that "it takes a Nvidia M40 GPU 14 days to finish just one 90-epoch ResNet-50 training execution on the ImageNet-1k dataset". Transfer learning makes it possible to train useful models with just a fraction of the original computing effort and with only a small amount of training data. Off-the-shelf pre-trained machine learning models have gained academic interest, especially after the ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC 2014) [16], which provided a noticeable leap in Open ML model performance used for image classification.

2.5. Raw and Tuned models

Off-the-shelf open ML models have been used as templates for modification and changes that may alter the resource cost, performance, and accuracy of models in the same task. In this paper, we use the term *raw model* for a reused pre-trained ML model from an open-source provider. We use the term *tuned models* for those pre-trained ML models that have had their parameters manually tuned/changed, their structure modified, or that have been amalgamated together for the same task. Amalgamations may consist of structural merging or even majority voting between multiple models of the same type trained with different subsets of the same retraining dataset.

2.6. Related Studies

We are not familiar with other systematically conducted reviews directly related to shared Open ML usage. The closest related study was done by Nguyen *et al.* [21] in the form of an expertly opinioned and observed survey. It provides information regarding different statistically popular Open ML models, frameworks, and hardware. In addition, there are also lists of open-source ML libraries, such as the one curated by The Institute for Ethical Machine Learning, available at GitHub [22].

3. Research approach

The systematic mapping study as a form of a systematic review is a well-defined research method to identify, analyze, and synthesize all relevant studies regarding a particular research question or topic area [23, 3]. The systematic mapping study method was chosen for this paper because it aims at a holistic, credible, and fair overview of studies on shared pre-trained Open ML model usage.

3.1. Protocol

An important step when performing a systematic review is the development of a protocol. The protocol specifies all steps performed during the review, increasing its rigor and reliability. The protocol was constructed following the systematic review guidelines [24]. The protocol used in this study was also inspired and adapted from the procedure introduced by Mahdavi-Hezavehi *et al.* [25] in their review.

The procedures start with the research question definition, search strategy identification, and search scope selection. After that, study inclusion and exclusion criteria were formed based on the research questions. An empirical data extraction form was created based on the research questions. The data collection was conducted by filling out the data extraction form from the analyzed studies found and included in searches.

3.2. Research questions

This study covers the following research questions:

- RQ1: What solutions are used for shared pre-trained Open ML models?
- RQ2: How does research compare different open ML models, datasets, and frameworks?
- RQ3: What evidence is available on the performance and evolvability of pre-trained Open ML model solutions?

Table 1
Electronic sources searched, and the numbers of papers found and finally included.

Electronic sources	Number of hits per search (duplicates)	Number of selected results per search
SpringerLink Journals	854 (76)	392
IEEE Xplore	233 (0)	107
Total	1087 (76)	499

3.3. Search strategy

The automatic search was conducted by executing search strings on search engines of the following digital libraries:

- IEEE Xplore: ("machine learning" OR "Deep learning") AND ("pre-trained model" OR "pre-trained models")
- SpringerLink: ("machine learning" OR "Deep learning") AND ("pre-trained model" OR "pre-trained models")

The following study inclusion criteria were used for the inclusion of the papers:

- I1: The paper experiments with the usage of pre-trained ML models. Experiments are required to collect information in order to analyse solutions, adoptions, and evolvability.

The following study exclusion criteria were used:

- E1: The paper does not feature the usage of pre-trained Open ML models. If the focus of a paper was on other than Open ML models, the paper was excluded.
- E2: The Open ML model used in the paper is presented as a novel one. The model is not yet shared if it is novel.
- E3: Paper is an editorial, technical report, position paper, abstract, keynote, opinion, tutorial summary, panel discussion, or a book chapter.
- E4: Paper is grey literature. Grey literature is argued to be of lower quality than papers published in journals and conferences as they usually are not thoroughly peer-reviewed [26].

The numbers of papers found and included during the search phase are shown in Table 1. The publication date of searched papers was limited between 2013-2020. On SpringerLink, the search was limited to journal articles due to a hign number of conference papers (over 1900) requiring identification. We decided to prefer journals

Table 2
Data extraction form.

#	Field	Reason
F1	Author(s)	Overview
F2	Year	Overview
F3	Title	Overview
F4	Keywords	Overview
F5	Citation count (From Google Scholar[1])	Overview
F6	Used models (Table 6)	RQ1, RQ2
F7	Used datasets (Table 7)	RQ1, RQ2
F8	Used frameworks (Table 8)	RQ1, RQ2
F9	Application domain (Table 4)	Overview
F10	Evidence level (Table 3)	Overview
F11	Raw Model Effectiveness	RQ3
F12	Tuned Model Effectiveness	RQ3

[1] https://scholar.google.fi/

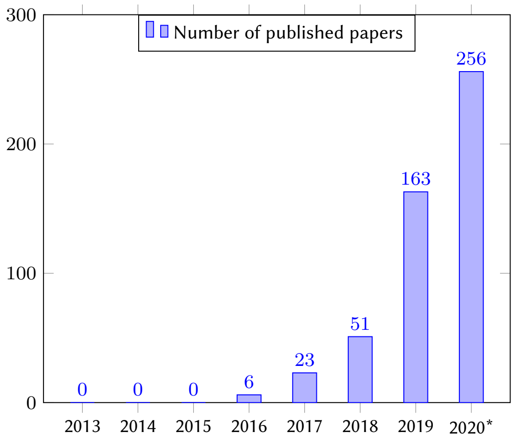

Figure 1: Papers per year. 2020 consists of papers published until the end of October.

Table 3
The number of papers at the evidence levels.

Evidence levels	Number of papers (%)
1 (No evidence)	0
2 (Demos)	2 (0,4%)
3 (Expert opinions, observations)	0
4 (Academic studies)	488 (97,8%)
5 (Industrial studies)	9 (1,8%)
6 (Industrial evidence)	0

Table 4
Domains addressed by studies.

Domain	Number of papers (%)	Domain	Number of papers (%)
Biology	120 (24,0%)	Industrial	2
Medical	97 (19,4%)	Space Engineering	1 (0,2%)
Transportation	13 (2,6%)	Chemistry	1
Geography	6 (1,2%)	Meteorology	1
Surveillance	6	Geology	1
Music	3 (0,6%)		
Business	2 (0,4%)		
Electrical Engineering	2		
Astronomy	2		

Table 5
ML tasks addressed by studies.

Task	Number of papers (%)	Task	Number of papers (%)
Image classification	395 (79,2%)	Face recognition	13 (2,6%)
		Image translation	9
Image object detection	159 (31,9%)	Text detection	7 (1,4%)
		Speech recognition	4 (0,8%)
Text classification	48 (9,5%)	Data mining	5 (1,0%)
Video object detection	14 (2,8%)	Music generation	3 (0,6%)
		Texture categorization	1 (0,2%)
Video classification	13 (2,6%)	Human activity recognition	1

over conference papers for their more meticulous peer-review process compared to the shorter review time of conference papers. IEEE consisted of only 41 journal articles in total, so we decided to include conference papers in order to provide a more comprehensive sample.

3.4. Data Extraction

Data was extracted using the data extraction form (Table 2). For the evidence levels (F10), the classification system proposed by Alves *et al.* [27] was used consisting of six levels:

- 1. No evidence.
- 2. Evidence obtained from demonstration or working out toy examples.
- 3. Evidence obtained from expert opinions or observations.
- 4. Evidence obtained from academic studies (e.g., controlled lab experiments).
- 5. Evidence obtained from industrial studies (i.e., studies are done in industrial environments, e.g., causal case studies).
- 6. Evidence obtained from industrial application (i.e., actual use of a method in industry)

4. Results and analysis

This section first gives an overview of the identified studies and extracted information. After that, the research questions are answered by representing the extracted data and summarizing the data as an answer to each question.

4.1. Results overview and demographics

After performing the search and selection described above in Section 3, we included 499 papers in the data analysis.

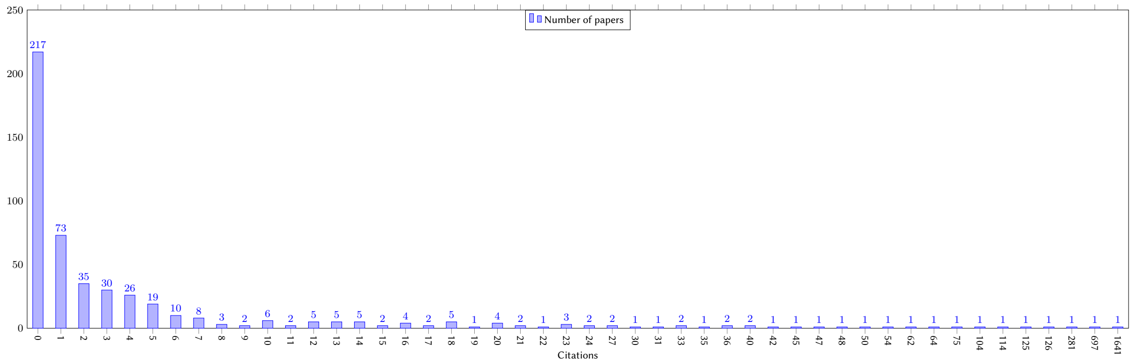

Figure 2: Study citation counts.

The number of papers published each year between January 2013 and October 2020 is shown in Figure 1. The first papers started to appear only in 2016, and the highest number of studies was published in 2020. Figure 1 also indicates incremental interest following the ILSVRC 2014 competition taking into account writing and publishing delays of over a year. In particular, the increase in interest has been exponential rather than linear over recent years.

The number of papers at different evidence levels is shown in Table 3. Almost all papers (97.8%, i.e., 488 papers) provided Level 4 evidence (academic studies) of their findings. The few remaining papers present Level 2 (Demonstration) or Level 5 (industrial study) evidence. As this review focused on existing Open ML models, it is unsurprising that at least Level 4 is achieved. However, there needs to be more practice-oriented studies at Levels 5 and 6. All studies provide mostly academic or industrial-level research, but most do not offer enough comparative evidence to adopt their used models. Only a few studies critically examined the potential influence of different actors, such as the researchers' bias, sponsors, and the quality of tests used to validate their study.

Figure 2 shows the citation counts for studies. As can be seen, the lowest and highest citation counts are 0 and 1641, respectively. 464 (around 93.0%) have a citation count in the range of 0–20, and 35 papers (7.0%) have high citation counts in the range of 21–126. A few significant outliers were S111 (review of deep learning for time series classification), S203 (new simple approach for batch normalization), and S31 (new edge detection algorithm), with citation counts of 281, 697, and 1641, respectively.

The domains and tasks addressed by studies are shown in Tables 4 and 5. 79.2% of the studies (i.e., 395 papers) addressed Image classification, while the second most popular class was Image object detection with 31.9% of studies (i.e., 159 papers). Biology was addressed by 24.0% of the studies. Medical got addressed by 19.4% (i.e., 97 papers) of the studies. Text classification was addressed by 9.5% of the studies. The rest of the domains had less than 15 studies addressing them. Some studies addressed more than one domain, so the total number of papers in the table is more than the amount reviewed. In summary, the domains related to images or videos are clearly the most prevalent.

4.2. RQ1: Open ML models, datasets, and frameworks

To answer this research question, the data of F6 (used models), F7 (used datasets), and F8 (used frameworks) were analyzed from the data extraction form and summarized in what follows. Because some studies used more than one model in their comparisons, the total numbers of papers are more than the amount reviewed.

Table 6 presents pre-trained Open ML models used by the studies and the number of studies applying each pre-trained model. 149 different Open ML models were identified. The most popular pre-trained Open ML model, i.e., VGG-16, is used by 168 (33.7%) studies, while the next most popular AlexNet and ResNet-50 are included in 100 (20.2%) and 99 (19.8%) studies, respectively. The majority of the models have less than six studies using them.

Table 7 shows datasets used to train and test in the studies. A total of 49 different datasets were identified, and in 114 studies, the framework was not specified. ImageNet is used by 60.7% of the studies (i.e., 303 papers). In contrast, the second most popular MS COCO and Google News Word2Vec are included in a significantly smaller number of studies, i.e., 19 studies each. The majority of the datasets have less than three studies using them. 114 studies do not mention their datasets explicitly, and thus, they could not be extracted.

Table 8 lists the frameworks used in the studies. We identified in total 37 different frameworks, and the framework is not specified in 153 studies. 21.8% of the studies

Table 6
Pre-trained Open ML Models used by studies.

Model	Number of papers (%)	Model	Number of papers (%)	Model	Number of papers (%)	Model	Number of papers (%)
VGG-16	168 (33,7%)	Inception-ResNet-V2	22 (4,4%)	ResNet-18	13 (2,6%)	DenseNet201	9
AlexNet	100 (20,2%)	Word2Vec	19 (3,8%)	BERT	13	Mask RCNN	9
ResNet-50	99 (19,8%)	ResNet-152	18 (3,6%)	CaffeNet	13	DenseNet121	9
Inception-V3	77 (15,2%)	Xception	17 (3,4%)	VGG-F	11 (2,2%)	U-Net	8 (1,6%)
VGG-19	64 (12,8%)	MobileNet	17	VGGNet	11	YOLOv2	8
GoogleNet	41 (8,2%)	MobileNetV2	16 (3,2%)	SqueezeNet	10 (2,0%)	ZFNet	7 (1,4%)
Faster R-CNN	39 (7,8%)	GloVe	15 (3,0%)	ResNet	9 (1,8%)	VGG-Face	6 (1,2%)
ResNet-101	32 (6,4%)	YOLOv3	14 (2,8%)	Inception	9	**118 Other models**	1 - 5 (0,2 - 1,0%)

Table 7
Datasets used by studies.

Dataset	Number of papers (%)	Dataset	Number of papers (%)	Dataset	Number of papers (%)	Dataset	Number of papers (%)	Dataset	Number of papers (%)
ImageNet	303 (60,7%)	Kinetics	3 (0,6%)	MUSE	1	Places2	1	FER-2013	1
MS COCO	19 (3,8%)	CASIA-WebFace	2 (0,4%)	Tweet2Vec	1	Yang 91	1	SFEW 2.0	1
GoogleNews Word2Vec	19			Universal Sentence Encoder	1	BSD 300	1	ELMo	1
		Places	2			DIV2K	1	HAM10000	1
GloVe	15 (3,0%)	MNIST	2			B200	1	DeepSpeech	1
BERT	13 (2,6%)	XLNet	2	Change-Detection.net	1	G200	1	Youtube-8M	1
PASCAL	10 (2,0%)	Conceptnet Numberbatch	2			Flickr	1	Calamari	1
VGG-Face	7 (1,4%)			FCVID	1	MS-Celeb-1M	1	BiGRU	1
FastText	6 (1,2%)	R-Net	1 (0,2%)	CIFAR-100	1	CelebFaces	1	*Unspecified*	114 (22,8%)
Darknet53	6 (0,8%)	TheoryTab	1	Tesseract	1	VGG-Face2	1		
PlaceS362	4 (0,8%)	GRID	1	Ocropy	1	VGG16.V2. CalimeMod	1		
CIFAR-10	4	Sports 1M	1	VOC2007	1				

(i.e., 109 studies) used the TensorFlow framework. The second most often used is Keras, with 16.2% studies (i.e., 81 studies). Over a half of the frameworks have less than four studies using them. 30.7% of studies did not explicitly mention their frameworks and were unresolvable.

4.2.1. Summary to RQ1

There is a large diversity in shared Open ML model usage across studies, although VGG-16 stood out as the most popular. Many different Open ML models have been used in different studies. A large part of the variety appears due to the multiple application domains addressed, as seen in Table 4, and their requirements for models with specialized application domain capabilities, such as word relation recognition (Parsey McParseface, word2vec) and music recognition (MidiNet, etc.). With frameworks and especially datasets, we see less disparity between the studies. Especially in the case of datasets, although there is a larger number of different datasets than OpenML models, ImageNet stands out as dominant, appearing in 303 (60.7%) studies, and most datasets appear in at most a couple of studies. This lack of disparity may be due to the lack of interest and usefulness regarding the less widely appearing datasets or the significant time and work effort required to train new models based on them.

Table 8
Frameworks used by studies.

Framework	Number of papers (%)	Framework	Number of papers (%)	Framework	Number of papers (%)	Framework	Number of papers (%)	Framework	Number of papers (%)
TensorFlow	109 (21,8%)	OpenCV	10 (2,0%)	Conceptnet Numberbatch	2	R-Net	1	Calamari	1
Keras	81 (16,2%)	Darknet	9 (1,8%)			DeepSpeech	1	ULMfit	1
Caffe	78 (15,6%)	BERT	8 (1,6%)	Theano	2	BioBERT	1	AdverTorch	1
MATLAB	33 (6,5%)	fastText	6 (1,2%)	Magenta	1 (0,2%)	ELMo	1	Orange	1
PyTorch	28 (5,6%)	Google Colab	4 (0,8%)	MidiNet	1	MUSE	1	WEKA	1
Word2Vec	19 (3,8%)	Tesseract	3 (0,6%)	SyntaxNet	1	Universal Sentence Encoder	1	BiGRU	1
GloVe	17 (3,4%)	Doc2Vec	2 (0,4%)	Ocropy	1			Tweet2Vec	1
MatConvNet	15 (3,0%)	XLNet	2	AraVec	1			*Unspecified*	153 (30,7%)

Table 9
The number of Open ML Models in the studies.

Count	Number of papers (%)
1	202 (40,4%)
2	119 (23,8%)
3	87 (17,4%)
4	34 (6,7%)
5	20 (4,0%)
6	17 (3,6%)
7	6 (1,2%)
8	6 (1,2%)
9	3 (0,6%)
10	1 (0,2%)
11	2 (0,4%)
13	1 (0,2%)
14	2 (0,4%)
22	1 (0,2%)

4.3. RQ2: Comparisons of open ML models, datasets, and frameworks

In addition to the data analyzed in the above RQ1, the number of different Open ML models that were compared within each study was analysed in order to form a conceivable adoption preference based on the comparison. In Table 9, the number of Open ML models which are explicitly compared by studies is listed. Many studies that do not compare Open ML models with other open models, however, make a comparison to differently licensed ones, such as copyrighted or proprietary models and frameworks without openly available source code. In total, 59.6% of studies compare their results to other Open ML models, and 35.8% compare to more than one Open ML model. 40.4% of the papers do not compare their results to other Open ML models.

Likewise, an analysis was carried out on how many different datasets were compared in each study. Table 10 lists the number of datasets that were compared in studies. 10.4% of the studies (i.e., 52 papers) compare the use of other datasets, and 15 of them to more than one. As also seen in Table 10, at least 66.1% of the papers do not compare their results to other datasets. Due to the unresolved datasets used by 22.8% of the studies, an unspecified category was added to the table.

Finally, Table 11 lists the number of frameworks compared in studies. 15.0% of the studies (i.e., 75 studies) compare their results with other frameworks, and 13 of them to more than one. As also visible in Table 11, at least 54.5% of the studies do not compare their results with other frameworks. The results are not very accurate due to the unresolved framework datasets used by nearly a third (30.5%) of the studies. These were categorized as unspecified in the table.

Table 10
The number of datasets.

Count	Number of papers (%)
1	330 (66,1%)
2	37 (7,4%)
3	6 (1,2%)
4	8 (1,6%)
7	1 (0,2%)
Unspecified	114 (22,8%)

Table 11
The number of frameworks in the studies.

Count	Number of papers (%)
1	272 (54,5%)
2	62 (12,4%)
3	10 (2,0%)
4	3 (0,6%)
Unspecified	152 (30,5%)

4.3.1. Summary to RQ2

The number of Open ML models used per study was counted to see how well they are represented between comparisons. Around six out of ten studies compared Open ML models to other Open ML models, which can be considered quite a large amount.

Also, the amount of dataset and framework comparisons were counted, but they provided significantly less instrumental results. Only one-tenth of studies compared the results of different datasets and about slightly over one-tenth with different frameworks. Those that did compare datasets and frameworks were also mainly limited to only comparing two. However, unlike Open ML Models, datasets and frameworks have only a few dominant designs that are widely applied (cf. RQ1). It should also be taken into account that dataset and framework results are inaccurate because many studies do not explicitly mention what was used by name.

The limited amount of contrastive studies did not offer enough information for reliable results of domain-specific or general Open ML model adoptions. The scale of dataset and framework adoption is also unclear. However, also scientific literature provides some evidence on the popular adoption rates of certain Open ML models: VGG-16 and AlexNet, frameworks such as TensorFlow, and datasets like ImageNet.

Table 12
Model type performance measured by studies.

Models measured	Number of papers (%)
Raw Model Only	268 (53,7%)
Tuned Model Only	110 (22,0%)
Both Models	121 (24,2%)

4.4. RQ3: Evidence available on the performance and evolvability of pre-trained Open ML model solutions

The model type used for performance measurement studies is shown in Table 12. 272 studies (53.8%) give results for raw model performance that only use transfer learning. 112 studies (22.1%) provided results for models that were ensembled, modified, or tuned by researchers. Tuning ranges from individual value changes to layer replacement. 122 (24.1%) studies provided both raw and tuned performance results. The performance measurement results provided by the studies are not directly comparable and thus not listed.

Table 13 shows the performance of pre-trained Open ML Model solutions described in the studies when the studies compare their results to the previous state-of-the-art solutions, such as NN, SVM, and constant feature classification algorithms. Improvement in Table 13 consists of studies describing at least one of the Open ML Models outperforming the state-of-the-art solutions. The partial improvement consists of studies describing Open ML Models being competitive and partially outperforming in specific categories, such as lower computational cost without much performance disadvantage compared to others. Decline consists of studies where Open ML models performed worse than other solutions. As seen from Table 13 majority of studies, 74.1% (i.e., 370 papers), provide minor to moderate improvement compared to previous methods. 24 (4.8%) studies found a decline compared to the state-of-the-art performance when using Open ML models. 108 (21.0%) studies do not show improvement, lack comparison, or have mixed results when using Open ML models.

4.4.1. Summary to RQ3

In total, 24.2% of studies do not provide results for raw models to compare the effectiveness of their tuned models. This can cause the lack of ability for readers to know if the tuned model is better than the raw model. Still, Table 12 cannot be directly used to evaluate the trustworthiness of an evolved model, and comparisons can be made to other evolved models, such as presented in the highest cited S31. Studies also have diverse evaluation approaches to evaluate the raw and tuned models. The studies could not directly be used to evaluate the evolvability of models, but gave positive promise for their evolvability.

Table 13
Performance range of Open ML Model solutions described in studies when compared to previous state of the art.

Perfomance	Number of papers (%)
Improvement	308 (61,7%)
Partial improvement	62 (12,4%)
No improvement/ Mixed results/ Lack of comparison	105 (21,0%)
Decline	24 (4,8%)

5. Discussion

This section summarizes and discusses the main findings, limitations to the review, and threats to validity.

5.1. Main findings

Most studies either carry out an academically novel test with an Open ML model in a specific domain or show the results from performing major novel tuning, modifying the model, or using an amalgamation of models. The main focus of novel modifications in the reviewed studies lies in showing that what is found is overall better than the previous, not as much on rigorous comparison of models, datasets, and frameworks, and discussion of these options for the specific use case. Researchers' and publishers' bias toward positive novel discoveries and the lack of showing failed experiments may have a negative effect on public opinion about the trustworthiness of scientific studies after real-world results contradict them [28].

The majority of the Open ML models found in the studies appear to be DNN models used on image classification tasks that have participated in the ILSVRC competition [16]. A considerably large share of studies uses Open models on TensorFlow framework with ImageNet as the dataset. While some designs can become dominant for specific domains or purposes, the lack of diversity in used frameworks and datasets in academic studies can negatively affect the use of more diverse options used in the industry. For datasets, there might eventually come a point when other datasets cannot compete with that one large dataset that has been poured with the majority of interest and available data.

The wide use of the same frameworks and models pre-trained with the same datasets may also increase the probability of somebody knowing and eventually exploiting faults in certain pre-trained models used in future industrial applications, for example, through physical-world attacks on sensors [29]. Mitigation could be made possible by obscuring known vulnerabilities by forming multiple variants of models trained on different subsets of the commonly used dataset or by adding randomized statistical and mathematical initial insights that get mostly turned over during training but leave a unique mark on the model.

Using pre-trained models also raises ethical questions about the data used to train the models, along with ownership of the models. Where did the data come from? Who owns it? Does it contain sensitive information? Is the data comprehensive enough to be fair and unbiased? Is the data traceable or exploitable? Careless use weakens the explainability of the model and its potential errors, and the developer of the newly-trained model easily loses ownership of the model if they are not careful. While the developer is usually considered responsible for the model, problems or exploits originating in the original training data can be difficult to defend, excuse, and, most importantly, fix.

5.2. Research direction for future work

An exponential increase and a very large number of papers already in 2020 make it clear that future research must be focused. As academic studies are prevalent, industrial studies, including longitudinal behavior observations, are one direction for future research.

There was a lot of discrepancy in methods used by studies to evaluate and compare the usefulness of an ML model to other ML methods in the same domain. It would be beneficial for researchers to form a unified evaluation and comparison model to be used by studies for certain types of ML models under different domains to provide more easily comparable results.

It is also suggested that researchers increase rigorous result testing to other similar Open models in their studies. Including results from different combinations of frameworks and datasets would provide more comprehensive results but may cause a lot of work and redundancy. Though redundancy in studies has been discovered to be effective in cancelling out researcher bias [28], it could still be avoided by using an open curated database of results for researchers to refer to, like openml.org [30] for example.

5.3. Reflection of review

To objectively evaluate a systematic review, Kitchenham et al. proposed four quality questions for systematic reviews [27]:

Are inclusion and exclusion criteria described? It is considered that this review meets this criterion as it explicitly defined and explained inclusion and exclusion criteria in Section 3.4.

Is the literature search likely to have covered all relevant studies? This criterion is not fully met as we searched only two digital libraries and did not apply other strategies, such as snowballing or manual searches. The second weakness is that our search string was limited and did not cover different synonyms of the terms. The third weakness is that we did not search grey literature and, in particular, arxiv.org, which publishes especially several relevant pre-prints for the topic. While aware of these weaknesses, the corpus of included papers covered was quite large (499 studies) and can be considered a somewhat representative sample of all research.

Did the reviewers assess the quality/validity of the included studies? It is considered that the quality and validity of studies were not assessed well enough due to the lack of data used to analyze them. The evidence levels and citation counts are insufficient to determine quality and validity conclusively. The limitation to peer-reviewed published papers was used to set a threshold for the quality. However, this study is a mapping study where quality assessment is not quintessential.

Were the basic data/studies adequately described? One of the main limitations of systematic reviews is the inaccuracy in data extraction that we also encountered. There were difficulties in extracting relevant information from selected studies. Several studies do not, for example, explicitly mention in which domains their modifications to a model could be used. This could cause the researcher's interpretation bias to affect the final extracted data. The issue was mitigated by listing only generic domains, which caused the high count of generalizing results that might not actually implicate the true domain of a study. Another problem while analyzing frameworks and pre-train datasets used by studies was that papers do not always clearly mention what was used. Finally, due to a large number of studies, we can present only summary information in this paper, leaving full details to supporting online material.

6. Conclusion

The main goal of this study was to investigate how pre-trained Open ML models, frameworks, and datasets are shared and used in different domains through a systematic mapping study research method. Based on a relatively large sample, the reviewed 499 studies provide a listing of Open ML models, frameworks, and datasets used in research as well as their relative popularity. The studies consist of many different domains, which saw benefits ranging from minor decline to moderate improvement compared to previously used machine learning methods. This indicated that pre-trained Open ML models and frameworks show positive promise for improvement in machine learning. Most of the models in the studies were used under the TensorFlow framework with ImageNet as the pre-train dataset. Most studies were academic, and only a few industrial studies were identified. More industrial-level studies are required to be reviewed in order to have more reliable and accurate representations of Open ML model performance in the real world.

Suggestion for the future is to increase the coverage of studies and modify the review inclusion criteria for study extraction when assessing the usage of shared pre-trained Open ML models. Another more conclusive option is to prototype and create a constantly updated open curated database of results for different Open ML models running on different frameworks using different pre-train datasets. These configurations would then be run and tested on different domain datasets. The different possible combinations and results could then be calculated on a cloud platform, with the only requirement of having to insert the new model, dataset, or framework addition through a curated application form.

Acknowledgements

This work was partly funded by Business Finland under IML4E and IVVES of ITEA programme.

References

[1] A. Serban, K. van der Blom, H. Hoos, J. Visser, Adoption and effects of software engineering best practices in machine learning, in: Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 2020, pp. 1–12.

[2] Z.-H. Zhou, Learnware: on the future of machine learning., Frontiers Comput. Sci. 10 (2016) 589–590.

[3] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, in: 12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12, 2008, pp. 1–10.

[4] S. Russell, P. Norvig, Artificial intelligence: a modern approach (2002).

[5] R. O. Duda, P. E. Hart, D. G. Stork, Pattern classification, John Wiley & Sons, 2012.

[6] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.

[7] G. E. Hinton, T. J. Sejnowski, T. A. Poggio, et al., Unsupervised learning: foundations of neural computation, MIT press, 1999.

[8] B. Perens, et al., The open source definition, Open sources: voices from the open source revolution 1 (1999) 171–188.

[9] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 675–678.

[11] F. Chollet, et al., Keras, https://github.com/fchollet/keras, 2015.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (2009) 10–18.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[14] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL: https://www.tensorflow.org/, software available from tensorflow.org.

[15] A. Vedaldi, K. Lenc, Matconvnet: Convolutional neural networks for matlab, in: Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 689–692.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[18] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[19] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.

[20] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, K. Keutzer, Imagenet training in minutes, in: Proceedings of

the 47th International Conference on Parallel Processing, 2018, pp. 1–10.

[21] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. L. García, I. Heredia, P. Malík, L. Hluchỳ, Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey, Artificial Intelligence Review 52 (2019) 77–124.

[22] Curated list of open source libraries, https://github.com/EthicalML/awesome-production-machine-learning/, 2020. [Online; accessed 13-04-2020].

[23] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, 2007.

[24] S. Keele, et al., Guidelines for performing systematic literature reviews in software engineering, Technical Report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.

[25] S. Mahdavi-Hezavehi, M. Galster, P. Avgeriou, Variability in quality attributes of service-based software systems: A systematic literature review, Information and Software Technology 55 (2013) 320–343.

[26] B. A. Kitchenham, P. Brereton, M. Turner, M. K. Niazi, S. Linkman, R. Pretorius, D. Budgen, Refining the systematic literature review process—two participant-observer case studies, Empirical Software Engineering 15 (2010) 618–653.

[27] V. Alves, N. Niu, C. Alves, G. Valença, Requirements engineering for software product lines: A systematic literature review, Information and Software Technology 52 (2010) 806–820.

[28] J. P. Ioannidis, Why most published research findings are false, PLos med 2 (2005) e124.

[29] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.

[30] J. Vanschoren, J. N. van Rijn, B. Bischl, L. Torgo, Openml: Networked science in machine learning, SIGKDD Explorations 15 (2013) 49–60.